Efficient training of multilayer perceptrons using principal component analysis

Bunzmann, C; Biehl, M; Urbanczik, R

Published in:
Physical Review E

DOI:
10.1103/PhysRevE.72.026117

Publication date:
2005

Efficient training of multilayer perceptrons using principal component analysis

Christoph Bunzmann,[1] Michael Biehl,[2] and Robert Urbanczik[1]
[1]*Institut für Theoretische Physik, Universität Würzburg Am Hubland, D-97074 Würzburg, Germany*
[2]*Institute for Mathematics and Computing Science, University Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands*
(Received 15 March 2005; published 16 August 2005)

A training algorithm for multilayer perceptrons is discussed and studied in detail, which relates to the technique of principal component analysis. The latter is performed with respect to a correlation matrix computed from the example inputs and their target outputs. Typical properties of the training procedure are investigated by means of a statistical physics analysis in models of learning regression and classification tasks. We demonstrate that the procedure requires by far fewer examples for good generalization than traditional online training. For networks with a large number of hidden units we derive the training prescription which achieves, within our model, the optimal generalization behavior.

DOI: 10.1103/PhysRevE.72.026117 PACS number(s): 84.35.+i, 89.90.+n, 64.60.−i, 87.19.La

I. INTRODUCTION

Multilayered neural networks can serve as parametrizations of classification or regression schemes [1,2]. Their ability to approximate any reasonable input-output relation to arbitrary precision is achieved by interconnecting sufficiently many identical, simple processing units. The most attractive feature of such devices is their ability to learn from example data.

In supervised learning, the choice of the network parameters or weights is based on a set of examples of the objective task. Frequently, the training procedures are guided by the optimization of appropriate cost functions which measure the network performance with respect to the available data. In regression problems, for instance, gradient based methods can be used to minimize the quadratic deviation of the network output from the target values, the so-called training error.

After training, the network implements a hypothesis or approximate realization of the unknown rule. Its successful application to novel data—which was not contained in the training set—has been termed generalization and it is the ultimate goal of supervised learning. In a continuous regression problem the generalization error would quantify, for instance, the expected quadratic deviation for a test input. In the context of classification, the generalization error corresponds to the probability of misclassifying novel data.

One key objective of the theory of learning is the evaluation of learning curves, i.e., the generalization error after training as a function of the number of examples in the exploited data set. In this context, methods from the statistical physics of disordered systems have been applied successfully in the analysis of model learning scenarios, see Refs. [1,2] for reviews. The key ingredients of the approach are (a) the consideration of large networks with many degrees of freedom, i.e., the thermodynamic limit and (b) the performance of averages over the stochastic nature of the training process as well as over the disorder contained in the training data. In doing so it is possible, for instance, to evaluate typical learning curves for a given network architecture, rule complexity, training procedure, and specified statistical properties of the example data.

Artificial neural networks are frequently assembled from many simple structures which repeat over the network. As a consequence, the occurrence of symmetries is practically inevitable. It has been demonstrated within a variety of learning situations that the properties of networks in the training phase may depend strongly on these symmetries [1,2]. In multilayered neural networks the output is often invariant under exchange or permutation of the network branches connecting to the so-called hidden units. This permutation symmetry has to be broken in the course of learning in order to yield good generalization behavior.

In offline or batch training the entire given set of example data is used in defining a cost function. The typical outcome of training can be studied in the framework of equilibrium statistical physics by treating the weights as thermodynamic variables related to a formal energy which corresponds to the cost function [1,2]. Accordingly, a temperaturelike parameter controls the typical energy or tolerated training error. In such a setting, the symmetries can lead to the emergence of phase transitions, e.g., in a discontinuous drop of the generalization error at a critical size of the training set [1,2].

Another important model scenario is the widely investigated online training of continuous feedforward neural networks [3], for instance, by use of stochastic gradient descent [4]. Here, one finds pronounced plateau states in the learning curves, see, e.g., Refs. [4–7] for further references. Successful learning requires hidden unit specialization, i.e., the breaking of the permutation symmetry. This effect can be delayed significantly, if no *a priori* knowledge about the target rule is available.

For large networks trained from N-dimensional randomized i.i.d. example inputs, a statistical physics analysis of online gradient descent has been done for various learning situations. These investigations show that, without initial specialization, successful training beyond the plateau states is impossible if the number of examples grows linearly with N, i.e., with the number of adjustable parameters [4,6]. Note that also optimized training schedules or modified gradient procedures as studied in, e.g., Refs. [8–10] require initial nonzero specialization. Analogous effects have been demonstrated in the context of classification problems for

the training of networks with binary units, see, e.g., Refs. [1,2,11–13].

The question arises whether these findings reflect a genuine difficulty in the training of multilayer networks or just result from the use of inappropriate training schemes. A major purpose of this work is to demonstrate that the latter is the case. To this end we put forward and investigate in detail a recently proposed alternative approach to the supervised training of multilayered networks which is applicable to regression as well as classification problems [14].

The key idea of the procedure is to effectively reduce the dimensionality of the learning problem in a first step. It corresponds to performing a principal component analysis (PCA) with respect to an appropriately chosen correlation matrix of the example data. In the second phase of training, the necessary specialization can then be achieved by adaptation of a few parameters, the number of which increases only quadratically with the number K of hidden units in the system.

In order to demonstrate the potential usefulness of the suggested training scheme, we investigate its typical properties in model scenarios of regression and classification. We demonstrate that in both cases good generalization is achievable if the number of examples is only linear in N, without the requirement of *a priori* specialization. The algorithm retains the most attractive feature of online training, as the example data are not memorized explicitly and storage needs do not grow with their number.

The paper is organized as follows. In Sec. II we outline the model situation, present our algorithm, and discuss its basic ideas. In Secs. III and IV we present the theoretical analysis of the two phases of training. Results are discussed and compared with Monte Carlo simulations of the training process. The theory is extended to networks with a large number of hidden units in Sec. V. In Sec. VI we demonstrate how variational methods can be used to optimize the training prescription with respect to the generalization behavior. Finally we summarize and conclude by discussing open questions and potential follow-up projects.

II. IDEA AND ALGORITHM

As a prototype multilayer architecture, we consider a committee machine (CM) with K hidden units. Its output τ is defined by the activation functions $h(x)$ in the hidden layer, $g(x)$ in the output layer, and by the K parameter vectors $B_i, i=1,2,\ldots,K$:

$$\tau(B^T\xi) = g(f_0(B^T\xi)), \tag{1}$$

$$f_0(B^T\xi) = K^{-1/2}\sum_{i=1}^{K} h(B_i^T\xi), \tag{2}$$

where $\xi \in \mathbb{R}^N$ is the input vector and $B=[B_1,\ldots,B_K]$ is the $N\times K$ matrix of the unknown parameter vectors $B_i \in \mathbb{R}^N$. In the following, we will restrict our analysis to the case of orthonormal vectors $B_i^T B_j = \delta_{ij}$.

The activation function of the hidden units is assumed to be odd, $h(-x)=-h(x)$, and bounded. Whenever numbers are given, they refer to a sigmoidal of the form $h(x)=l_\lambda \mathrm{erf}(\lambda x)$ with $\lambda > 0$.

We will consider two variations of the above basic architecture which are suitable for classification and regression problems, respectively. The two cases differ in the choice of the output activation function $g(x)$.

(i) *Regression, soft committee machine*. An invertible function $g(x)$ realizes continuous network outputs. Here we will concentrate on the case $g(x)=x$ and use the term soft committee machine (SCM) for the corresponding network.

(ii) *Classification, hard committee machine*. The choice $g(x)=\mathrm{sgn}(x)$ corresponds to a binary classification of input data, we will refer to the architecture as the hard committee machine (HCM). Here, the limiting case $h(x)=\mathrm{sgn}(x)$ for the sigmoidal hidden unit activation represents a network with only binary units.

Clearly, the rule complexity that can be considered within our restricted model is limited. Nevertheless we expect the setting to capture the essential features of supervised learning in multilayered networks to a large extent. Modifying, for instance, the SCM to become a universal approximator requires only one additional adaptive parameter per hidden unit and the relaxation of the assumption that $B_i^T B_j = \delta_{ij}$ [15]. This would complicate the analysis in the following without providing further insight into the problems addressed here.

The ultimate goal of learning is to estimate the parameter vectors B_i, which we shall refer to as teacher vectors, from a training set $\mathcal{P}$ of P examples $[\xi^\mu, \tau(B^T\xi^\mu)]$ of the input-output relationship. A good estimate $J=[J_1,\ldots,J_K]$ will result in a small generalization error $\epsilon_g(J)$.

Throughout the following we assume that all input data consist of independently drawn random components with zero mean and unit variance. Writing $\langle\cdots\rangle_\xi$ for the average over the corresponding probability distribution of an input ξ, this is

$$\epsilon_g(J) = \begin{cases} \dfrac{1}{2}\langle[\tau(B^T\xi) - \tau(J^T\xi)]^2\rangle_\xi, & \text{(invertible } g\text{)}, \\ \dfrac{1}{2}\langle 1 - \tau(B^T\xi)\tau(J^T\xi)\rangle_\xi, & [g=\mathrm{sgn}(x)]. \end{cases} \tag{3}$$

Note that for HCM networks with $g(x)=\mathrm{sgn}(x), \tau=\pm 1$, the generalization error ϵ_g indeed gives the probability for the misclassification of a random input since $[1-\tau(B^T\xi)\tau(J^T\xi)]/2 \in \{0,1\}$.

Throughout the following, the prefactor l_λ in the hidden unit activation $h(B_i^T\xi)=l_\lambda \mathrm{erf}(\lambda B_i^T\xi)$ will be taken to satisfy the constraint $\langle h^2(B_i^T\xi)\rangle_\xi = 1$. Standard online methods of learning aim at improving the student performance by iterative modifications of the network parameters, each based on only a single example. Typically, they display two different relevant scales for the number of examples on which learning proceeds: A single, common weight vector $W \in \mathbb{R}^N$ shared by all hidden units is quickly obtained from a number of examples linear in N. It corresponds to a plateau state with an intermediate value of ϵ_g [4–6]. Further improvement of ϵ_g requires specialization of the hidden units, i.e., the breaking of permutation symmetry. Without *a priori* knowledge, i.e.,

initial specialization, the plateau can only be left after presenting a much larger number of examples which is superlinear in N. Results published in, e.g., Refs. [4,6] suggest that the required number of examples should grow as $N \ln N$, instead.

Several methods have been devised to improve the success of training in this context. These include sophisticated learning rate schedules, prescriptions using second order derivatives of the error measure, or other concepts that have proven useful in offline training. These methods can reduce the length of plateaus significantly when starting with little initial specialization, see, e.g., Refs. [8–10]. However, the problem remains that without initial knowledge, successful specialization and good generalization require a number of examples which is much larger than the dimension KN of the parameter space.

In the following we will present a method which, indeed, allows for successful training from a number of examples that grows linearly with the number of network weights. Hence, we will consider training sets of size $P=\alpha KN$ with the accordingly rescaled number of examples α.

The very idea of our method is to decrease the dimensionality of the problem in a first step. A principal component analysis (PCA) of a properly defined $N\times N$ matrix C^P reduces the task to an optimization problem in only K^2 dimensions. The matrix C^P is computed from the outer products of the input vectors ξ weighted by a function of the known teacher output. It may furthermore depend on the field $W^T\xi$ of an auxiliary perceptron with weights $W \in \mathbb{R}^N$.

In an iterative formulation of the scheme, the update of matrix C^μ and auxiliary weight vector W^μ upon presentation of example $\mu+1$ is written as

$$W^{\mu+1} = (\mu+1)^{-1}[\mu W^\mu + \tau(B^T\xi^{\mu+1})\xi^{\mu+1}], \tag{4}$$

$$C^{\mu+1} = (\mu+1)^{-1}\left[\mu C^\mu + \xi^{\mu+1}\xi^{(\mu+1)T} \times F\left(\frac{W^{\mu T}}{|W^\mu|}\xi^{\mu+1}, f_0(B^T\xi^{\mu+1})\right)\right]. \tag{5}$$

While the choice of Hebbian learning, Eq. (4), for the estimation of the auxiliary perceptron vector W might be replaced with more sophisticated estimates, we restrict our analysis here to this particularly simple case in the following. The weight function F in Eq. (5) has to be specified in order to define a particular training algorithm. Its choice is in the center of the following analysis.

In regression problems, i.e., for $g(y)=y$, a particularly simple choice for the weight function is $F(x,y)=-y^2$, which does not make use of the auxiliary weight vector. This corresponds, in a sense, to an extension of Hebbian learning, taking into account correlations of the terms $\tau\xi$

Since only the value of τ, as defined in Eq. (1), is available to the training algorithm, the specific transfer function g can restrict the possible dependence of $F(x,y)$ on y, as, for instance, in the case of the HCM with $g(y)=\mathrm{sgn}(y)$. As we will demonstrate in the following section, one simple and successful choice of the weight function for classification with an HCM is $F(x,y)=\Theta(-xy)$. It compares the student output with that of the auxiliary perceptron and puts emphasis on examples where they disagree.

The limiting case $P\to\infty$ reveals how the spectrum and eigenvectors of C^P are related to the unknown teacher vectors, and how this relation is influenced by the choice of the weight function F. Assuming that the components ξ_i^μ are independent Gaussian random variables with zero mean and unit variance, it is straightforward to analyze the spectrum of C^P for $P\to\infty$. In this limit the auxiliary perceptron weights satisfy the condition $W^P \propto B_{\mathrm{av}}$ with

$$B_{\mathrm{av}} = K^{-1/2}\sum_{i=1}^{K} B_i \tag{6}$$

and

$$\frac{(W^P)^T\xi}{|W^P|} = B_{\mathrm{av}}^T\xi = f_{\mathrm{av}}(B^T\xi). \tag{7}$$

The eigenvalues of C^P are found easily, as—for the theoretical analysis—we can choose a coordinate system that simplifies the problem very much. The examples are split into the teacher space components y_i and orthonormal components $z_i, i=1,2,\ldots,N-K$. Thus it is easily seen that the orthonormal space is an eigenspace with eigenvalue λ_0. Due to the permutation symmetry of the hidden units, in the teacher space one symmetric eigenvector B_{av} with eigenvalue $\bar{\lambda}$ and an $(K-1)$-dimensional eigenspace spanned by B_1-B_j $(j=2,\ldots,K)$ with eigenvalue λ_Δ can be identified:

$$\lambda_0 = \langle z_1^2\rangle_z\langle F(f_{\mathrm{av}}(y),f_0(y))\rangle_y = \langle F(f_{\mathrm{av}}(y),f_0(y))\rangle_y,$$

$$\bar{\lambda} = \langle [y_1+(K-1)y_2]y_1F(f_{\mathrm{av}}(y),f_0(y))\rangle_y,$$

$$\lambda_\Delta = \langle (y_1-y_2)y_1F(f_{\mathrm{av}}(y),f_0(y))\rangle_y, \tag{8}$$

where $y\in\mathbb{R}^K$ is the vector of normally distributed random variables y_i and the short-hand $f_{\mathrm{av}}(y)$ is defined in Eq. (7). B_{av} itself is of little interest since for large P one also has $W^P\propto B_{\mathrm{av}}$, and it is thus simpler to use Hebb's rule (4). The eigenspace of λ_Δ is the one we wish to identify by PCA.

Here the importance of properly selecting the weight function F is evident: λ_Δ should be a distinct, extremal eigenvalue, so that its eigenspace can be found by PCA. Numerically, it is easiest to compute the relevant eigenspace if λ_Δ is the largest eigenvalue. Hence, we shall only consider such choices for F.

For a finite number P of training examples the degeneracy in the spectrum is broken by random fluctuations. Nevertheless, a computation of the $N\times(K-1)$-dimensional matrix of the eigenvectors

$$\Delta = [\Delta_1^P,\ldots,\Delta_{K-1}^P] \tag{9}$$

of C^P associated with the largest eigenvalues, yields an empirical estimate of the space spanned by the difference vectors B_1-B_j. Together with W^P, this yields an estimate of the teacher space. Training a given network is thus reduced to finding a $K\times K$ matrix Γ, such that the set of student vectors $J=[W^P,\Delta]\Gamma$ minimizes $\epsilon_g(J)$. We want to optimize the de-

tection of the teacher space by choosing F carefully. The ultimate aim is to lower the minimal $\epsilon_g(J)$ which can be obtained by optimizing Γ in the second training phase.

In the following, we address central aspects of the procedure from a theoretical point of view and test the predictions in Monte Carlo simulations of the learning process. First, the overlap of the $K-1$ extremal eigenvectors with the teacher space is calculated for matrices C^P defined by general architectures (g,h) of the teacher, see Eq. (1). It is measured by the *subspace overlap*

$$\rho = (K-1)^{-1/2}\mathrm{Tr}(\Delta B^T B \Delta^T)^{1/2}, \tag{10}$$

where Δ is defined in Eq. (9). This is a sensible measure because ρ is invariant with respect to orthonormal reparametrizations of Δ and it attains its maximal value of 1 if and only if the Δ_j^P lie in the space spanned by the B_i. Hence, ρ can be interpreted as the cosine of the "angle" between the subspace spanned by the Δ_j^P and the one spanned by the teacher vectors B_i.

Next, we obtain a prediction for the achievable generalization error given the subspace overlap ρ. Furthermore, in the limiting case of $K\to\infty$, the equation for ρ is solved analytically. This enables us to determine the optimal weight function F with respect to the typical generalization behavior in large networks, see Sec. VI.

III. TEACHER SPACE VIA PCA: THEORETICAL ANALYSIS

A good weight function F will result in a large subspace overlap ρ obtained from a given number of examples P. In order to identify the optimal weight function for a given architecture, we need to calculate the typical value of ρ for training sets with P examples for general F, first.

The following calculations rely on the thermodynamic limit $N\to\infty$, i.e., large input layers. This allows us to introduce order parameters and evaluate the corresponding partition function and free energy in terms of a saddle point integration. We choose the number μ of examples growing linearly in KN and introduce $\hat{\mu}$ as the corresponding rescaled quantity $\mu=\hat{\mu}KN$.

In the following, we restrict the analysis to weight functions that yield an estimate of the teacher space via the eigenspace of the $K-1$ largest eigenvalues. If a given F produced such an estimate via the eigenspace of the $K-1$ smallest eigenvalues, we could apply the transformation $F\to -F$.

The eigenspace of the largest eigenvalue of C^P can be found by maximizing $X^T C^P X$ with respect to the N-dimensional vector X of length 1. Hence, we consider the partition function

$$Z = \int dX \exp(\beta P X^T C^P X), \tag{11}$$

where the integration is over the unit sphere in $\mathbb{R}^N$. For large N the typical properties of the maximization problem are found by calculating the training set average $\langle \ln Z\rangle_{\mathcal{P}}$ and taking the limit $\beta\to\infty$. The replica trick is used to evaluate the average of the logarithm via $\lim_{n\to 0}\partial_n\langle Z^n\rangle_{\mathcal{P}}$ and we will proceed assuming replica symmetry. The disorder averaged partition function Z^n of the n-fold replicated system reads

$$\langle Z^n\rangle_{\mathcal{P}} = \left\langle \int dX^{[n]} H(\mathcal{P},X^{[n]}) \right\rangle_{\mathcal{P}}, \tag{12}$$

where the integration is over n copies of the unit sphere and

$$H(\mathcal{P},X^{[n]}) = \prod_{\mu=1}^{P}\prod_{a=1}^{n} e^{\beta\gamma^{\mu,a}(X^{aT}\xi^\mu)^2}$$

with

$$\gamma^{\mu,a} \equiv F\left(\frac{W^{\mu-1^T}\xi^\mu}{|W^{\mu-1}|}, f_0(B^T\xi^\mu)\right).$$

In contrast to many related problems, the Gibbs weight $H(\mathcal{P},X^{[n]})$ does not factorize over the examples, here. This is due to the dependence of $\gamma^{\mu,a}$ on $W^{\mu-1}$. So, in a first step, we rewrite the training set average as

$$\langle H(\mathcal{P},X^{[n]})\rangle_{\mathcal{P}} = \left\langle \left\langle e^{\beta\Sigma_{a=1}^n \gamma^{P,a}(X^{aT}\xi^P)^2}\right\rangle_{\xi^P} H(\mathcal{P}',X^{[n]})\right\rangle_{\mathcal{P}'},$$

where $\mathcal{P}'$ is the training set with the last pattern removed. The ξ^P average depends on the X^a via their mutual overlaps and via their overlaps with the B_i and with W^{P-1}. Among the latter, the dynamic overlaps $X^{aT}W^{P-1}$ are rather troublesome from a technical point of view. But W^{P-1} is estimating B_{av}, whereas a vector picked from the Gibbs density, see Eq. (11), estimates an eigenvector orthogonal to B_{av}. So, it is hardly conceivable that the Gibbs density is concentrated on W^{P-1} in the thermodynamic limit just due to random fluctuations. Hence, we assume and exploit that $X^{aT}W^{P-1}=0$.

A second point is, that the average over ξ^P depends on the overlaps between the teacher vectors and W^{P-1}. Due to this fact, it depends on all of the previous examples, even if $X^{aT}W^{P-1}=0$, and the Gibbs weight still does not factorize for finite N. But $B_i^T W^{P-1}$ is self-averaging for $N\to\infty$, and so the averages over ξ_P and $\mathcal{P}'$ do decouple in this limit.

We thus obtain, in the thermodynamic limit, the ratio of the average Gibbs weight for P and $P-1$ patterns as

$$E(\alpha,X^{[n]}) = \frac{\langle H(\mathcal{P},X^{[n]})\rangle_{\mathcal{P}}}{\langle H(\mathcal{P}',X^{[n]})\rangle_{\mathcal{P}'}} = \left\langle \prod_{a=1}^{n} e^{\beta(X^{aT}\xi^P)^2 F(r(\alpha)B_{\mathrm{av}}^T\xi^P+\sqrt{1-r(\alpha)^2}\nu, f_0(B^T\xi^P))}\right\rangle_{\xi^P,\nu},$$

where ν is a zero mean, unit variance Gaussian and independent of ξ^P. Further, $r(\alpha)$ is the overlap $B_i^T W^{P-1}$ which for large N is of the form

$$r(\alpha) = [1 + c_{g,h}/(K\alpha)]^{-1/2}, \tag{13}$$

where the coefficient $c_{g,h}$ depends only on the activation functions g and h.

In summary we obtain

$$\langle Z^n\rangle_{\mathcal{P}} = \int dX^{[n]} \exp\left(KN \int_0^\alpha d\hat{\mu} \ln E(\hat{\mu}, X^{[n]}) \right). \quad (14)$$

We have thus recovered a standard situation, since $E(\hat{\mu}, X^{[n]})$ depends on $X^{[n]}$ only via the overlaps $X^{aT}X^b$ and $X^{aT}B_i$.

Now, by the usual arguments, detailed in the Appendix, a replica symmetric parametrization allows us to calculate $M(\alpha)$, the typical value of $\max_X N^{-1} X^T C^P X$ for large N:

$$M(\alpha) = \max_R \min_\chi \frac{1 - R^T R}{2\chi} + \alpha K \mathcal{G}(\chi, 1 - R^T R + (R^T y)^2). \quad (15)$$

Our main interest is to obtain the value of the K-dimensional vector R which maximizes Eq. (15). This gives the typical overlap $B^T X$ of the vector maximizing $X^T C^P X$ with the teacher space. Further, the functional $\mathcal{G}$ used in Eq. (15) is defined by an average over an isotropic K-dimensional Gaussian y with zero mean and unit variance components as

$$\mathcal{G}(\chi, \phi(y)) = \frac{1}{\alpha} \int_0^\alpha d\hat{\mu} \langle \phi(y) \langle G(y_p(\hat{\mu}), f_0(y)) \rangle_\nu \rangle_y$$

$$G(y_p(\hat{\mu}), f_0(y)) = \frac{F(y_p(\hat{\mu}), f_0(y))}{1 - 2\chi F(y_p(\hat{\mu}), f_0(y))},$$

$$y_p(\hat{\mu}) = r(\hat{\mu}) f_{\mathrm{av}}(y) + \sqrt{1 - r(\hat{\mu})^2}\, \nu. \quad (16)$$

Since Eq. (15) is a quadratic form in R, we rewrite it as

$$M(\alpha) = \max_R \min_\chi R^T A(\chi) R + a(\chi), \quad (17)$$

where $a(\chi) = 1/2\chi + \alpha K \mathcal{G}(\chi, 1)$ and the K by K matrix $A(\chi)$ has elements

$$A_{ij}(\chi) = \alpha K \mathcal{G}(\chi, y_i y_j) - \delta_{ij} \left(\frac{1}{2\chi} + \alpha K \mathcal{G}(\chi, 1) \right).$$

To further analyze Eq. (17), assume we have found the solution $\chi(R)$ of the minimization in χ for a given choice of R. We then need to maximize $R^T A[\chi(R)]R + a[\chi(R)]$. The gradient with respect to R of this function is simply $2A[\chi(R)]R$, since $\chi(R)$ is stationary. So the maximization problem can only have a solution $R \neq 0$ if $A[\chi(R)]$ is singular and such a solution is an eigenvector of $A[\chi(R)]$ with eigenvalue 0.

The eigenvectors of A do not depend on χ. One eigenvector, with eigenvalue $A_{11} + (K-1)A_{12}$, is $\Sigma_{i=1}^K e_i$, where $e_1, \ldots, e_K$ is the standard basis of $\mathbb{R}^K$. Other eigenvectors are $e_1 - e_j$ $(j = 2, \ldots, K)$ and these have the eigenvalue $A_{11} - A_{12}$, with degeneracy $K-1$.

If the weight function F is chosen properly, $A_{11} - A_{12}$ is the larger of the two eigenvalues and defines a space, where Eq. (17) is indeed maximal. Due to its degeneracy, our analysis of the properties of the single vector X maximizing $X^T C^P X$ in fact yields the properties of the $K-1$ eigenvectors of C_P with the largest eigenvalues in the thermodynamic limit. Also due to degeneracy, we may parametrize Eq. (17) by setting $R = \rho(e_1 - e_j)/\sqrt{2}$ to obtain an extremal problem in only two variables ρ and χ. Note that ρ is then indeed the subspace overlap introduced in Eq. (10). Now, one easily sees that if Eq. (17) yields a nonzero ρ, the solution satisfies

$$0 = A_{11}(\chi) - A_{12}(\chi),$$

$$\rho^2 = -\frac{a'(\chi)}{A'_{11}(\chi) - A'_{12}(\chi)},$$

where the prime denotes the differentiation with respect to χ. For later use, we note that a more explicit version of these equations is

$$0 = \alpha K \mathcal{G}(\chi, (y_1 - y_2)^2 - 2) - \frac{1}{\chi}, \quad (18)$$

$$\rho^2 = \frac{1 - 2\chi^2 \alpha K \mathcal{G}'(\chi, 1)}{1 + 2\chi^2 \alpha K \mathcal{G}'(\chi, (y_1 - y_2)^2 - 2)}. \quad (19)$$

An explicit solution is possible if $\mathcal{G}(\chi, \cdots)$ does not depend on the number of examples presented. On the one hand, this is true if the weight function is of the form $F(y_p(\hat{\mu}), f_0(y)) = f(g(f_0(y)))$. For the soft committee machine such weight functions can be found, indeed, e.g., $f(x) = -x^2$ or $f(x) = -x^2 + 1$, which is a minor modification of the simple analogy to Hebbian learning. Note, however, that in cases where a good estimate of the perceptron vector is available such choices of f are inferior to the more general ansatz which includes a dependence of F on $W^T \xi$.

In order to test the theoretical predictions for the subspace overlap ρ we have performed Monte Carlo simulations for, both, SCM and HCM with $K = 3$ three hidden units. Here, we have chosen the following weight functions.

(i) *Regression, SCM with* $g(x) = x, h(x) = l_1 \mathrm{erf}(x)$. The choice $F(x, y) = 1 - y^2$ is a minor modification of the example discussed in Sec. II, but improves the result considerably.

(ii) *Classification, HCM with* $g(x) = h(x) = \mathrm{sgn}(x)$. The weight function $F(x, y) = \Theta(-xy)$ corresponds to selecting examples on which the committee machine and the auxiliary perceptron disagree. Precisely these examples should contain considerable information about the internal parameters of the committee machine.

Simulation results for finite N are shown in Fig. 1 for the SCM and in Fig. 2 for the HCM, respectively. In both cases they are in good agreement with the theoretical prediction in the limit $N \to \infty$ for the subspace overlap ρ. Its dependence on the rescaled number of examples $\alpha = P/(KN)$ displays a second order phase transition from $\rho = 0$ at small α to nonzero values for $\alpha > \alpha_c$. This behavior is reminiscent of the results found in Ref. [16] for unsupervised principal component analysis of structured data. Note that the critical value α_c for the HCM is considerably larger than for the SCM. This agrees with the expectation that training a network with only binary units should be harder than a regression problem, in general. A detailed discussion of the dependence of α_c on the weight function F will be given in Secs. V and VI.

IV. FROM PCA TO THE STUDENT NETWORK

Student vectors are constructed by linear combinations of the vector W estimated via Eq. (4) and the eigenspace Δ

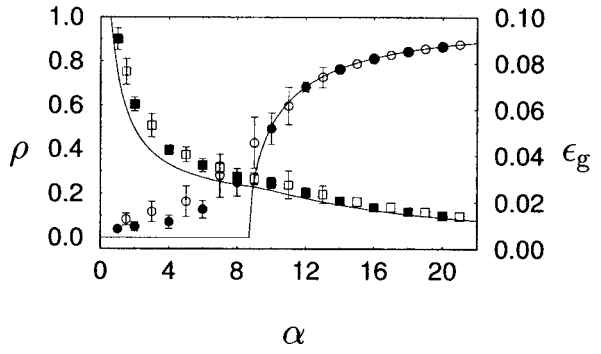

FIG. 1. Regression with an SCM: results for $K=3$ hidden units and activation functions $g(x)=x, h(x)=l_1\mathrm{erf}(x)$. The increasing curve displays the evolution of the subspace overlap ρ (left axis) with the rescaled number of examples $\alpha=P/(KN)$. The decreasing curve corresponds to the generalization error ϵ_g (right axis) as obtained in the second stage of our procedure. For ϵ_g, the value of α refers to the total number of examples used in both stages, $\alpha=(P+\hat{P})/(KN)$. Solid lines show the theoretical predictions for the thermodynamic limit $N\to\infty$. Symbols mark the results of simulations with $N=400$ (open symbols) and $N=1600$ (filled symbols), on average over five independent runs. Where not shown, error bars are smaller than the symbol size.

$=[\Delta_1^P,\ldots,\Delta_{K-1}^P]$ of the $K-1$ largest eigenvalues of the matrix given in Eq. (5). The K^2 optimization parameters are organized into a $K\times K$ matrix Γ, thus the student vector is $J=v\Gamma, v=[W,\Delta]$. The aim is to minimize the generalization error (3). In contrast to the usual online learning, we fit only a finite number of parameters, which lowers the risk of overfitting. To facilitate an exact analysis, we use $\hat{P}$ new examples in the training set and proceed as follows.

(i) *Regression, SCM with* $g(x)=x$ *in* Eq. (1). Here, simple on-line gradient descent can be chosen as an iterative optimization scheme. Given the example vector ξ and a teacher output τ the update is $\Gamma\to\Gamma+\delta\Gamma$ where $\delta\Gamma$ is a function of the fields of the hidden units $y=(v\Gamma)^T\xi$

$$\delta\Gamma_{ij}=\eta[\tau-f_0(y)]y_i h'(y_j). \tag{20}$$

(ii) *Classification, HCM with* $g(x)=\mathrm{sgn}(x)$ in Eq. (1). For the HCM we choose the update following [13]: Only if the outputs of teacher and student differ an update step is performed. Compared to Eq. (20) the derivative of the activation function of the hidden units is replaced by $\tau\exp(-sy_j^2)$, i.e., only those hidden units with fields close to zero on the scale set by the parameter $s^{-1/2}$ are updated considerably, and the direction is given by the teacher output

$$\delta\Gamma_{ij}=\eta\Theta(-\tau f_0(y))\tau y_i\exp(-sy_j^2). \tag{21}$$

For simulations, the learning rate η is chosen $\propto N^\gamma$, where $\gamma\in]-1,0[$ and $\hat{P}\propto PN^\kappa$ with $\kappa\in]-1-\gamma,0[$. In the limit $N\to\infty$ this means that $\eta\to 0$ but $\hat{P}\eta\to\infty$, since $P\propto N$. Thus, the gradient descent—for the invertible output unit of the SCM—turns into a deterministic search which is guaranteed to converge to a minimum of the generalization error. Note that with these scalings $\hat{P}/P\to 0$ for large N, so only a negligible number of additional training examples is needed for the online procedure in the second phase of training.

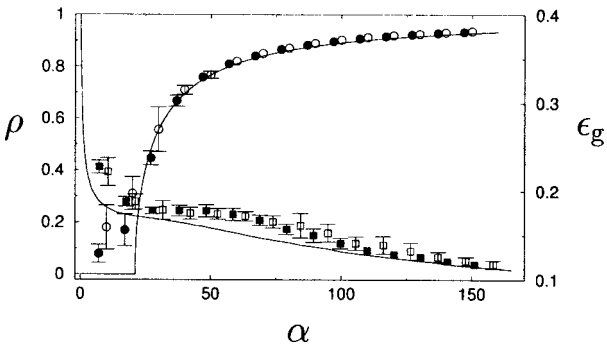

FIG. 2. Classification with an HCM: results for $K=3, g(x)=h(x)=\mathrm{sgn}(x)$, all other details as in Fig. 1. Note that significantly larger values of α are required for successful training than in the case of regression, cf. Fig. 1.

In Figs. 1 and 2, results for $K=3$ show how the increasing subspace overlap ρ does affect the generalization error. While for $\alpha<\alpha_c$ the generalization error drops due to an increasing overlap r of the perceptron vector, it does not show a plateau, but the steeply increasing ρ above α_c leads to a further decrease of the generalization error.

In Sec. III we presented the theoretical analysis of the subspace overlap ρ between the teacher space and the eigenspace obtained via PCA (10). It results in Eq. (13) for the typical overlap of the estimated averaged teacher vector W with B_{av} in Eq. (6) after $\hat{\mu}KN$ examples in the limit $N\to\infty$. Under the assumption that there exists a minimum of the generalization error which is unique up to permutations of the hidden units, the final state is one of minimal generalization error in the space spanned by $v=[W,\Delta]$. Thus, we are able to calculate the resulting value of ϵ_g after the gradient descent given the overlaps r and ρ by identifying the minimal generalization error.

In the limit $r\to 1, \rho\to 1$, the teacher space is perfectly described by the PCA and the Hebb vector, thus the minimal ϵ_g drops to 0. To obtain numbers for more general cases, we need to identify the minimal achievable value of ϵ_g when the matrix of student vectors is constrained to have the form $v\Gamma$. Each of the columns of v can be uniquely decomposed as $v_i=n_i+b_i$, where n_i is orthogonal to the subspace spanned by the teacher vectors and b_i is in this subspace. We assume that for large N, the n_i as well as the b_i are orthogonal to each other. While this could be checked for the PCA vectors by a replica analysis of maximizing $X^TC^PX+Y^TC^PY$ under the constraint $X^TY=0$, we have not done this, since the assumption seems plausible. Now, for any Γ, the assumption enables us to calculate the overlaps $B^Tv\Gamma$ given ρ and r, and thus the generalization error of the optimal student in the restricted space. So, also for the theoretical analysis, we have to solve the K^2-dimensional problem of finding an optimal Γ.

Since this becomes cumbersome for large K, we exploit the symmetries of the problem in the following manner. By construction, there exists an orthogonal matrix Γ_0 mapping the projections b_i onto the teacher vectors $B=[b_1,\ldots,b_K]\Gamma_0$. In a naive guess one might assume that an optimal student J

in the restricted space is given by $v\Gamma_0$, since nothing can be done about the noise vectors n_i. But since r and ρ are different in general, this need not be the case, and we use the ansatz $J=\kappa_1[(\kappa_2 W^P)\Delta_1^P\cdots\Delta_{K-1}^P]\Gamma_0$. Here, the parameters κ_1 and κ_2 allow us to control the length of the vectors in J and their correlation with $B_{\rm av}$. From κ_1,κ_2 the overlaps with the teacher space and each other are easily found as

$$R_{ij}=\frac{\kappa_1(\kappa_2 r-\rho)}{K}+\delta_{ij}\kappa_1\rho,\quad Q_{ij}=\frac{\kappa_1^2(\kappa_2^2-1)}{K}+\delta_{ij}\kappa_1^2$$

without requiring knowledge of Γ_0. Hence, for the theoretical analysis, we only need to solve a two-dimensional optimization problem.

To obtain ϵ_g from the order parameters, we need to specify the activation function of the hidden units h in Eq. (2). Given a linear activation of the output unit $g(x)=x$ in (1), we do not have to evaluate the integral over the $2K$ correlated fields in the hidden units numerically. It is solved analytically in Ref. [5], and we find the minimum on the two-dimensional manifold numerically.

For the hard output unit $g(x)=\mathrm{sgn}(x)$, we do not know of an analytic solution of the integral. We describe the fields in the student by a random variable for the part correlated with the teacher and K independent random variables. After the average over the K independent variables is done, the minimization involves in each step a $(K+1)$-dimensional integral.

The resulting learning curves $\epsilon_g(\alpha)$ for an SCM and HCM with $K=3$ hidden units are shown in Figs. 1 and 2, respectively. For comparison, the results of Monte Carlo simulations of the training process are displayed as well. The second order phase transition at α_c, which is clearly visible in $\rho(\alpha)$, persists in the generalization error $\epsilon_g(\alpha)$ as a kink in the learning curve which marks the onset of specialization.

V. LARGE K THEORY

In principle one can use the results in Sec. III to obtain the subspace overlap ρ for any given weight function F and number of hidden units K. However, the obtained equations are not very transparent. We thus exploit the major simplifications of the theory in the limiting case of many hidden units, i.e., $K\to\infty$ with $K\ll N$.

The first observation is that for large K the Hebb vector will be learned much faster than the subspace overlap ρ increases. The reason is that only on the order of N examples are needed for the former while a nontrivial result for ρ will require at least $\mathcal{O}(KN)$ examples. So, in this limit we can assume that $r=1$ already for small α and this simplifies the functional $\mathcal{G}[\chi,\phi(y)]$ in Eq. (16) considerably:

$$\mathcal{G}[\chi,\phi(y)]=\langle\phi(y)G(f_{\rm av}(y),f_0(y))\rangle_y, \tag{22}$$

where $f_{\rm av}(y)$ is defined in Eq. (7). To find the matrix coefficients which determine ρ in Eq. (19), we need to evaluate $\mathcal{G}(\chi,1)$ and $\mathcal{G}[\chi,(y_1-y_2)^2-2]$. For large K one is tempted to argue that $f_0(y)$ becomes Gaussian to simplify the average over the K fields $y_1,\ldots,y_K$ in Eq. (22). While this yields a useful result for $\mathcal{G}(\chi,1)$, for the second term one finds that $\mathcal{G}[\chi,(y_1-y_2)^2-2]=0$ for large K. This, in turn, would imply that in the large K limit only the trivial result $\rho=0$ could be obtained when using the PCA with αKN examples.

To determine the relevant scale of the learning curve, we need to evaluate the large K asymptotics of $\mathcal{G}[\chi,(y_1-y_2)^2-2]$ more precisely. Taylor expanding the y_1 and y_2 dependence of $G(f_{\rm av}(y),f_0(y))$ yields that up to $\mathcal{O}(K^{-3/2})$ corrections

$$\mathcal{G}[\chi,(y_1-y_2)^2-2]=\left\langle[(y_1-y_2)^2-2]\sum_{s=0}^{2}\sum_{i+j\le s}\frac{g_{i,j}}{i!\,j!}\times\frac{(y_1+y_2)^i[h(y_1)+h(y_2)]^j}{K^{(1/2)s}}\right\rangle_{y_1,y_2}, \tag{23}$$

where

$$g_{i,j}=\left\langle G^{(i,j)}\left(K^{-1/2}\sum_{k=3}^{K}y_k,K^{-1/2}\sum_{k=3}^{K}h(y_k)\right)\right\rangle_{y_3,\ldots,y_K}. \tag{24}$$

When carrying out the y_1 and y_2 average in Eq. (23) the constant and the linear terms vanish and so does the term with $i=2,j=0$. We are left with

$$\mathcal{G}[\chi,(y_1-y_2)^2-2]=K^{-1}(c_1g_{11}+c_2g_{02}),$$

where

$$c_1=\langle[(y_1-y_2)^2-2](y_1+y_2)[h(y_1)+h(y_2)]\rangle_{y_1,y_2}=2\langle(y_1^2-3)y_1h(y_1)\rangle_{y_1} \tag{25}$$

$$c_2=[\langle(y_1^2-1)h^2(y_1)\rangle_{y_1}-2\langle y_1h(y_1)\rangle_{y_1}^2]. \tag{26}$$

To determine the coefficients g_{11} and g_{02} from Eq. (24), we now argue that the joint density of $z_1=K^{-1/2}\sum_{k=3}^{K}y_k$ and $z'=K^{-1/2}\sum_{k=3}^{K}h(y_k)$ is Gaussian for large K, which reduces the calculations to two-dimensional integrals.

The stationarity condition (19) implies that $\chi\to 0$ with increasing K, since $\mathcal{G}'(\chi,1)$ stays finite. This considerably simplifies the coefficients (g_{11},g_{02}) and after some algebra we obtain that for $\alpha>\alpha_c(K)$

$$\rho^2=1-\frac{\alpha_c(K)}{\alpha}, \tag{27}$$

whereas $\rho=0$ for $\alpha<\alpha_c(K)$. The critical value $\alpha_c(K)$ is given in terms of averages over two independent zero mean and unit variance Gaussians z_1 and z_2:

$$\alpha_c(K)=\frac{4K\langle F(z_1,z')^2\rangle_{z_1,z_2}}{\langle c_1F^{(1,1)}(z_1,z')+c_2F^{(0,2)}(z_1,z')\rangle_{z_1,z_2}^2}$$

with

$$z'\equiv\gamma z_1+\sqrt{1-\gamma^2}z_2,\quad \gamma=\langle y_1h(y_1)\rangle_{y_1}, \tag{28}$$

where $F^{(m,n)}(\hat{x},\hat{y})=(\partial^m/\partial x^m)(\partial^n/\partial y^n)F(x,y)|_{x=\hat{x},y=\hat{y}}$.

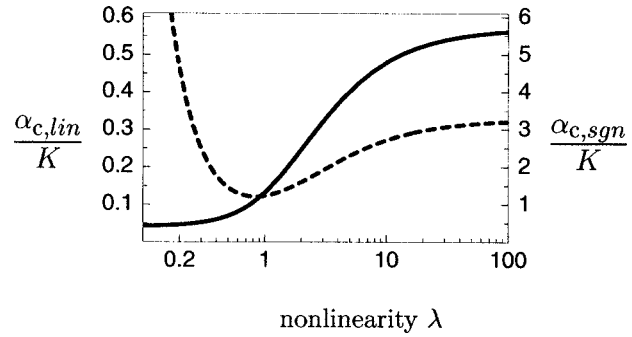

FIG. 3. The critical number of examples α_c as it results from the optimal weight function in the SCM with $g(x)=x$ ($\alpha_{c,\mathrm{lin}}$, solid line, left axis) and the HCM with $g(x)=\mathrm{sgn}(x)$ ($\alpha_{c,\mathrm{sgn}}$, dashed line, right axis). The graphs display the dependence on the nonlinearity λ in the hidden unit activation $h(x)=l_\lambda \mathrm{erf}(\lambda x)$.

VI. OPTIMAL WEIGHT FUNCTION

Equation (28) enables us to compute the optimal weight function in the limit $K\to\infty$, as well as the achievable generalization error and the minimal value of α_c.

To find the optimal weight function F, we first use partial integration for Gaussians $\langle f'(z_1)\rangle_{z_1}=\langle z_1 f(z_1)\rangle_{z_1}$, to eliminate the partial derivatives of F in Eq. (28). This yields

$$\alpha_c(K)=4K(1-\gamma^2)\frac{\langle F(z_1,z')^2\rangle_{z_1,z_2}}{\langle \phi(z_1,z_2)F(z1,z')\rangle^2_{z_1,z_2}}, \tag{29}$$

where $\phi(z_1,z_2)=c_1z_1z_2+[(c_2-c_1\gamma)/\sqrt{1-\gamma^2}](z_2^2-1)$.

It is important to recall how the activation of the output unit g limits the possible F. If the committee machine (1) uses $g(z)=\mathrm{sgn}(z)$, the field at the output unit z' can only be taken into account via its sign, while for invertible transfer functions at the output unit no restriction apply. Hence we have to treat the two cases separately in the following.

Case I: Regression with an SCM. Here, the Cauchy-Schwartz inequality implies that the optimal F satisfies $F(z_1,z')\propto\phi(z_1,z_2)$. In more explicit terms

$$F(x,y)=(c_2-c_1\gamma)\left(\frac{(y-\gamma x)^2}{1-\gamma^2}-1\right)+c_1x(y-\gamma x) \tag{30}$$

and the optimal value of α_c is

$$\alpha_{c,\min}=\frac{4K(1-\gamma^2)^2}{(1+\gamma^2)^2c_1^2-4\gamma c_1c_2+2c_2^2}. \tag{31}$$

For $h(x)=l_\lambda\mathrm{erf}(\lambda x)$, this can be evaluated analytically and the result is shown in Fig. 3. For the limiting cases of small and large λ we find

$$\frac{\alpha_{c,\min}}{K}=\frac{1}{24}+\frac{\lambda^2}{8}+\mathcal{O}(\lambda^4)\quad\text{for small }\lambda,$$

$$\frac{\alpha_{c,\min}}{K}=\left(\frac{\pi}{2}-1\right)-\frac{1}{\lambda}+\mathcal{O}(\lambda^{-2})\quad\text{for large }\lambda. \tag{32}$$

The result for small λ is remarkable since it predicts, that for the optimal F the relevant space is more and more efficiently detected, as the nonlinearity decreases. But in the limit $\lambda\to0$ the hidden unit activation function becomes linear, the teacher becomes a perceptron with weight vector B_{av} and the specialized overlap ρ must be zero. So the findings for the optimal F imply that the large K and small λ limit do not commute. This is not surprising, as the critical $\alpha_{c,\min}$, see Eq. (31), is undefined at $\lambda=0$, where both the numerator and the denominator are zero. So with λ decreasing to zero, larger and larger values of K are needed for our theory to hold.

While including the perceptron field in the weight function is not necessary for successful learning in the SCM, it can decrease the critical value α_c significantly. The optimal weight function under the restriction that it must not depend on x satisfies $F^{(1,1)}(x,y)=0$ in Eq. (28). In analogy to the optimization in the unrestricted case one obtains, then

$$F(x,y)=1-y^2, \tag{33}$$

$$\frac{\alpha_c}{K}=1/(8\lambda^4)+\mathcal{O}(\lambda^{-2})\quad\text{for small }\lambda \tag{34}$$

with the numerical example $\alpha_c=1.96$ K at $\lambda=1$ as compared to $\alpha_c=0.132$ K for the optimal weight function (30).

Case II: Classification with an HCM. If we want to minimize α_c under the restriction $g(x)=\mathrm{sgn}(x)$ of the teacher output (1), the weight function $F(z_1,z')$ may depend on z' only via $\mathrm{sgn}(z')$. Further, by the symmetries of the problem one can show that the optimal F satisfies $F(z_1,1)=F(-z_1,-1)$, i.e., F is of the form $F(z_1,z')=f(\mathrm{sgn}(z')z_1)$. It is convenient to split f in an even part f_+ and odd part f_-. Then the integration over z_2 in Eq. (29) is easily done. We get

$$\alpha_c=\frac{2K\pi(1-\gamma^2)^3}{(c_1-c_2\gamma)^2}\left\langle \exp\left(-\frac{1}{2}\frac{z_1^2}{\gamma'^2}\right)z_1f_-(z1)\right\rangle^{-1}_{z_1}$$
$$\times\left\langle f_-(z_1)^2+f_+(z_1)^2+2\,\mathrm{erf}\left(\frac{z_1}{\sqrt{2}\gamma'}\right)f_-(z_1)f_+(z_1)\right\rangle_{z_1}, \tag{35}$$

where $\gamma'=\sqrt{\gamma^{-2}-1}$. As f_+ appears only in the numerator, optimizing f_+ for a given f_- yields

$$f_+(z)=-\mathrm{erf}[(\sqrt{2}\gamma')^{-1}z_1]f_-(z_1). \tag{36}$$

Inserting this result in Eq. (35), one can again apply the Cauchy-Schwartz inequality to determine the optimal f_-. Finally, we find that the optimal F is $F(x,y)=f(x\,\mathrm{sgn}(y))$ with

$$f(z_1)=\frac{z_1\exp\left(-\frac{z_1^2}{2\gamma'^2}\right)}{1+\mathrm{erf}\left(\frac{z_1}{\sqrt{2}\gamma'}\right)}. \tag{37}$$

The lowest possible value of α_c for classification is thus

$$\alpha_{c,\min}=\frac{2K\pi(1-\gamma^2)^3}{(c_1-c_2\gamma)\langle\psi(z_1)^2\rangle_{z_1}}, \tag{38}$$

where

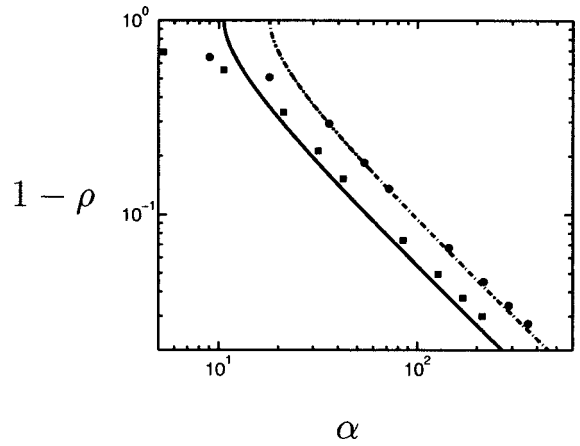

FIG. 4. Regression with an SCM with many hidden units and optimized weight function: subspace overlap ρ for $g(x)=x, h(x)=l_\lambda \mathrm{erf}(\lambda x)$. Theoretical prediction for $K=33, \lambda=3$ (solid line) and $K=33, \lambda=40$ (dashed line). Simulation results were averaged over five independent runs ($N=1200$) for $K=33, \lambda=3$ (circles), and $\lambda=40$ (squares). Standard error bars are smaller than the symbol size for all $\alpha>40$.

$$\psi(z_1) = \frac{z_1 \exp\left(-\frac{z_1^2}{2\gamma'^2}\right)}{\sqrt{1-\mathrm{erf}\left(\frac{z_1}{\sqrt{2\gamma'}}\right)}}.$$

For $h=l_\lambda \mathrm{erf}(\lambda x)$ the resulting α_c is shown in Fig. 3. The lowest values of $\alpha_{c,\min}$ are found for hidden units with $\lambda \approx 1$.

The theoretical predictions for the corresponding subspace overlap $\rho(\alpha)$ in SCM and HCM are shown in Figs. 4 and 5, respectively, for example values of the nonlinearity λ. Simulations with $K=33$ show good agreement for $\alpha>\alpha_c$, whereas deviations are larger in the vicinity of the critical value α_c as expected.

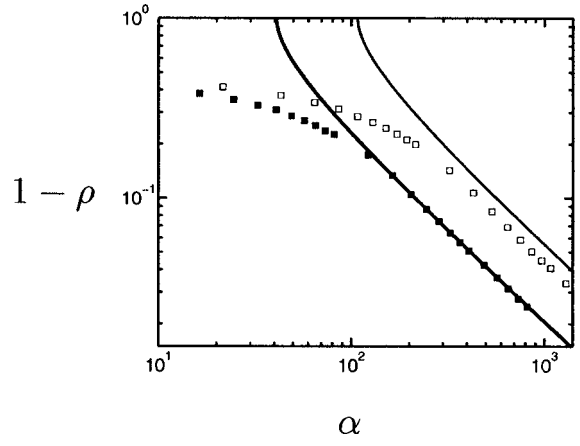

FIG. 5. Classification with an HCM with many hidden units and optimized weight function: subspace overlap ρ for $g(x)=\mathrm{sgn}(x), h(x)=l_\lambda \mathrm{erf}(\lambda x)$. Theoretical prediction for $K=33, \lambda=1$ (thick line) and $K=33, \lambda=100$ (thin line). Averages over five simulation runs ($N=450$) for $K=33, \lambda=1$ (filled symbols), and $\lambda=100$ (open symbols). Error bars are omitted for better readability.

VII. CONCLUSIONS

In summary, we have presented and discussed an efficient training algorithm for multilayer network architectures, typical properties of which can be analyzed exactly in the thermodynamic limit. Principal component analysis techniques are used to reduce the effective dimension of the learning problem in a first step. The necessary network specialization is then achieved in a lower-dimensional space of adaptive coefficients. We have shown that the basic idea of the algorithm can be applied in the context of, both, regression and classification problems.

Our results demonstrate that the algorithm is capable of yielding good generalization behavior for a number of training examples which is linear in the number of adaptive network parameters. In particular, *a priori* specialization is not required for the successful training of a given architecture. The analysis shows, furthermore, that the above features persist in the limit of infinitely many hidden units. Variational methods can be used to optimize the algorithms with respect to the expected, typical learning curves.

A comment is in place with respect to potential practical applications of the algorithm. We have assumed throughout the discussion that input data was generated according to an isotropic distribution. In practical applications one would clearly expect this assumption to be violated. However, it is always possible to *whiten* raw data by a linear transformation which yields a representation of the data with no structure on the level of second order statistics. Hereafter, our procedure can be applied.

We believe that our work should open new directions of research. Perhaps the most attractive feature of our prescription is that the number of hidden units need not be fixed prior to learning. The PCA procedure is performed without assumptions about the rule complexity. However, it provides information about the appropriate number K of hidden units, as we expect $K-1$ eigenvalues to separate from the rest of the spectrum. Hence, we suggest that our prescription could serve as a tool for, both, model selection and the actual training. Forthcoming investigations will focus on this aspect of the procedure.

ACKNOWLEDGMENTS

The work of two of the authors (C.B. and R.U.) was supported by the Deutsche Forschungsgemeinschaft.

APPENDIX: FREE ENERGY VIA REPLICA CALCULATION

In the thermodynamic limit, the integration over X_a is turned into an order parameter integration using $x_a=X_a^T\xi$ and the field with the teacher vector $y=B^T\xi$. The order parameters are

$$\langle x_a y_i \rangle_{x,y} = R_i^a = B_i X_a, \tag{A1}$$

$$\langle x_a x_b \rangle_x = Q^{ab} = X_a^T X_b. \tag{A2}$$

Rewriting the integration over the N-dimensional vectors as an integration over the order parameters is easily done.

This boils down to the evaluation of the integrand at the extremum

$$\begin{aligned}\lim_{N\to\infty}\frac{\langle \ln Z\rangle_{\mathcal{P}}}{N} &= \lim_{N\to\infty}\lim_{n\to 0}\partial_n\frac{\langle Z^n\rangle_{\mathcal{P}}}{N}\\ &= \lim_{N\to\infty}\lim_{n\to 0}\partial_n\int\prod_{a=1}^{n}\prod_{i=1}^{K}dR_i^a\prod_{b=1}^{n}dQ^{ab}\\ &\quad\times\exp\{N[\alpha K G_T(R,Q)+G_S(R,Q)]\}/N\\ &= \lim_{n\to 0}\partial_n\max_{R_i^a,Q^{ab}}\exp[\alpha K G_T(R,Q)+G_S(R,Q)],\end{aligned}$$

where $\alpha=P/(KN)$,

$$G_T=\frac{1}{\hat{\mu}}\int_0^{\alpha}d\hat{\mu}\times\ln\left(\left\langle\prod_{a=1}^{n}\exp[\beta x_a^2 F(y_p(\hat{\mu}),f_0(y))]\right\rangle_{x,y,\nu}\right), \tag{A3}$$

$$G_S=\frac{1}{N}\ln\left(\int dX^{(n)}\delta(R_i^a-X_a^T B_i)\,\delta(Q^{ab}-X_a^T X_b)\right). \tag{A4}$$

Following Ref. [12], we write the entropy term as a determinant of a matrix $G_S=\frac{1}{2}\ln\det(M)$, where

$$M=\begin{pmatrix}1 & \cdots & 0 & R_1^1 & \cdots & R_1^n\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ 0 & \cdots & 1 & R_K^1 & \cdots & R_K^n\\ R_1^1 & \cdots & R_K^1 & Q^{11} & \cdots & Q^{1n}\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ R_1^n & \cdots & R_K^n & Q^{n1} & \cdots & Q^{nn}\end{pmatrix}.$$

Restricting the maximization to the replica symmetric subspace of order parameters

$$Q^{ab}=\delta_{ab}+(1-\delta_{ab})q,$$

$$R_i^a=R_i \tag{A5}$$

and writing $R=(R_1,R_2,\ldots,R_K)$, we evaluate the determinant and get

$$\begin{aligned}G_S&=\frac{1}{2}\ln\det[\delta_{ab}+(1-\delta_{ab})q-R^TR]_{a,b=1,2,3,\ldots,n}\\ &=\frac{n}{2}\ln(1-q)+\frac{1}{2}\ln\left(1+n\frac{q-R^TR}{1-q}\right).\end{aligned} \tag{A6}$$

In the limit $n\to 0$, the derivative with respect to n yields

$$\lim_{n\to 0}\partial_n G_S=\frac{1}{2}\left(\ln(1-q)+\frac{q-R^TR}{1-q}\right). \tag{A7}$$

Evaluating the energy term, the fields x, which are correlated with y as given in Eq. (A1) are expressed via independent random variables

$$x_a=R^Ty+\sqrt{q-R^TR}z+\sqrt{1-q}z_a', \tag{A8}$$

where y,z,z' are independent Gaussian quantities with zero mean and unit variance. Having performed the average over the replica dependent variables z', we obtain an nth power in the argument of the logarithm instead of a product over n terms:

$$\begin{aligned}G_T=\frac{n}{\alpha}\int_0^{\alpha}d\hat{\mu}\ln\Bigg\langle&\exp\left(\frac{\beta F(y_p(\hat{\mu}),f_0(y))x_1'^2}{1-2\beta(1-q)F(y_p(\hat{\mu}),f_0(y))}\right)\\ &\times[1-2\beta(1-q)F(y_p(\hat{\mu}),f_0(y))]^{-1/2}\Bigg\rangle_{y,z,\nu},\end{aligned} \tag{A9}$$

where $x_1'=R^Ty+\sqrt{(q-R^TR)}z$. Now the derivative with respect to n at $n=0$ is taken, and the average over z is performed, to obtain the result for the energy term

$$\begin{aligned}\lim_{n\to 0}\partial_n G_T=\frac{1}{\alpha}\int_0^{\alpha}d\hat{\mu}\frac{1}{2}\Bigg\langle&\frac{2\beta F(y_p(\hat{\mu}),f_0(y))}{1-2\chi F(y_p(\hat{\mu}),f_0(y))}\\ &\times[q-R^TR+(R^Ty)^2]\\ &+\ln[1-2\chi F(y_p(\hat{\mu}),f_0(y))]\Bigg\rangle_{y,\nu},\end{aligned} \tag{A10}$$

where $\chi=\beta(1-q)$. We are interested in the limiting case $\beta\to\infty$, and it turns out the limiting process leaving χ fixed balances entropy and energy terms correctly. Given a fixed overlap R with the teacher, the vectors in the different replicas tend to be parallel as $\beta\to\infty$. The leading order terms of Eqs. (A7) and (A10) yield

$$\lim_{\beta\to\infty}\lim_{n\to 0}\partial_n\frac{G_T}{\beta}=\mathcal{G}[\chi,1-R^TR+(R^Ty)^2], \tag{A11}$$

$$\lim_{\beta\to\infty}\lim_{n\to 0}\partial_n\frac{G_S}{\beta}=\frac{1-R^TR}{2\chi}, \tag{A12}$$

where $\mathcal{G}(\chi,\cdots)$ is defined in the text. Together, we obtain

$$\frac{\langle \ln Z\rangle_{\mathcal{P}}}{\beta N}=\max_R\min_{\chi}\left\{\frac{1-R^TR}{2\chi}+\alpha K\mathcal{G}[\chi,1-R^TR+(R^Ty)^2]\right\}, \tag{A13}$$

from which we proceed in Sec. III.

[1] A. Engel and C. van den Broeck, *Statistical Physics of Neural Networks* (Cambridge University Press, Cambridge, 2001).

[2] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[3] S. Amari, IEEE Trans. Electron. Comput. **16**, 299 (1967).

[4] M. Biehl and H. Schwarze, J. Phys. A **28**, 643 (1995).

[5] D. Saad and S. A. Solla, Phys. Rev. Lett. **74**, 4337 (1995).

[6] M. Biehl, P. Riegler, and C. Wöhler. J. Phys. A **29**, 4769

(1996).
[7] *Online Learning in Neural Networks*, edited by D. Saad, (Cambridge University Press, Cambridge, 1998).
[8] D. Saad and M. Rattray, Phys. Rev. Lett. **79**, 2578 (1997).
[9] R. Vicente and N. Caticha, J. Phys. A **30**, L599 (1997).
[10] M. Rattray, D. Saad, and S. I. Amari, Phys. Rev. Lett. **81**, 5461 (1998).
[11] H. Schwarze and J. Hertz, Europhys. Lett. **21**, 785 (1993).
[12] M. Ahr and M. Biehl and R. Urbanczik, Eur. Phys. J. B **10**, 583 (1999).
[13] R. Urbanczik, Europhys. Lett. **35**, 553 (1996).
[14] C. Bunzmann, M. Biehl, and R. Urbanczik, Phys. Rev. Lett. **86**, 2166 (2001).
[15] G. Cybenko. Math. Control, Signals, Syst. **2**, 303 (1989).
[16] M. Biehl and A. Mietzner, Europhys. Lett. **24**, 421 (1993).